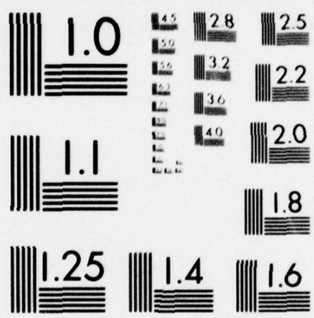
AD-A060 817 STANFORD UNIV CALIF SYSTEMS OPTIMIZATION LAB F/G 9/4
PROPERTIES OF CONJUGATE GRADIENT METHODS WITH INEXACT LINEAR SE--ETC(U)
JAN 78 L NAZARETH, J NOCEDAL N00014-75-C-0865
UNCLASSIFIED SOL-78-1 NL

1 OF 1
AD
A060817

END
DATE
FILMED
01-79
DDC


1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0

LEVEL II

12

Systems
Optimization
Laboratory

AD A060817

DDC FILE COPY

Department of Operations Research
Stanford University
Stanford, CA 94305

78 11 01 018

LEVEL II

(12)

SYSTEMS OPTIMIZATION LABORATORY
DEPARTMENT OF OPERATIONS RESEARCH
Stanford University
Stanford, California
94305

(6)

PROPERTIES OF CONJUGATE GRADIENT METHODS WITH INEXACT LINEAR SEARCHES

by

(10) L. Nazareth ~~and~~ J. Nocedal

(9) TECHNICAL REPORT. SOL 78-1

(11) January 1978

(12) 18p.

(14) SOL-78-1

(15) N00014-75-C-0865, EY-76-S-03-0326-PA-18

Research and reproduction of this report were partially supported by the Department of Energy Contract EY-76-S-03-0326 PA #18; the Office of Naval Research Contract N00014-75-C-0865; the National Science Foundation Grants MCS76-20019 and ENG77-06761.

408 765

PROPERTIES OF CONJUGATE GRADIENT METHODS WITH INEXACT LINEAR SEARCHES

by

L. Nazareth and J. Nocedal

1. Introduction

Conjugate gradient methods are extremely economical in their use of computer storage and yet surprisingly effective. They thus enjoy widespread popularity, in particular for optimizing non-linear objective functions involving a large number of variables, since the storage requirements of other methods may exceed that available. For example, in the MINOS code of Murtagh and Saunders [1], for optimizing a non-linear objective function subject to linear constraints (where the number of variables and constraints can both be very large), a suitable approximation to the optimum of an unconstrained non-linear objective function must be found during each iteration of the algorithm. When these unconstrained optimizations involve a sufficiently large number of variables, it is impractical to use a Quasi-Newton method. MINOS then switches to a conjugate gradient method.

Conjugate Gradient Methods were originally developed for solving systems of linear equations by Hestenes and Steifel [2] and subsequently applied to non-linear optimization by Fletcher and Reeves [3]. Many variants of the basic algorithm have since been suggested, e.g., [4], [5], [6], [7], [8], [9], [10], [11], [12], and extensive analysis carried out, e.g., [13], [14], [15], [16], and [17].

It is well known that when used to solve the linear system $Ax = b$, or equivalently to minimize the function $\psi(x) = a + b^T x + 1/2\, x^T Ax$, where A is an $n \times n$ positive definite symmetric matrix, the method of conjugate gradients can be looked upon as being a particular specialization of Gram-Schmidt orthogonalization of a given set of vectors in the inner product defined by A. Thus when used to minimize $\psi(x)$, successive approximations to the point where $\psi(x)$ attains its minimum are obtained by minimizing $\psi(x)$ in turn along each of a set of n search directions mutually orthogonal in the inner product defined by A i.e., mutually conjugate w.r.t. A. The search directions are not known beforehand. Instead, the search direction d_j at the current approximation to the minimum is developed from the gradient g_j at x_j and from conjugate search directions and gradients at previous iterations. Initially $d_1 = -g_1$. It can be shown that g_j is itself conjugate to $d_1, \ldots, d_{j-2}$. Any vector in the subspace spanned by g_j and d_{j-1} is therefore conjugate to $d_1, \ldots, d_{j-2}$. Thus we can simplify the Gram-Schmidt process for choosing a vector in the space spanned by g_j, $d_1, \ldots, d_{j-1}$ which is conjugate to $d_1, \ldots, d_{j-1}$, since the coefficients of components of this vector along $d_1, \ldots, d_{j-2}$ are zero, i.e., the Gram-Schmidt process specializes to choosing a d_j in the space spanned by d_{j-1} and g_j which is conjugate to d_{j-1}. Note that these statements, which underpin the method of conjugate gradients, require that line searches be exact.

In this paper we present some simple but fundamental results for the case when line searches are inexact. (Sections 4 and 5). These results again suggest methods of the conjugate gradient type which use very limited storage and which can be regarded as alternative specializations of the Gram-Schmidt orthogonalization process to the one discussed above (Section 6). These sections of the paper are preceeded by some introductory material in Sections 2 and 3.

2. Notation

(a) Search directions are denoted by d_j.

(b) Given a function $\phi(x)$ we shall denote its gradient at the point x_i by g_i.

(c) $y_i \triangleq g_{i+1} - g_i$.

(d) $\psi(x) = a + b^T x + 1/2\ x^T A x$ denotes a quadratic function. A is an $n \times n$ positive definite symmetric matrix.

(e) if $e_1, \ldots, e_t$ are a set of t n-vectors, $(c_1, \ldots, c_t)$ will denote the $n \times t$ matrix whose i-th column is c_i.

(f) x_{min} is the point where $\psi(x)$ attains its minimum.

3. Preliminaries

We shall appeal to the following two lemmas in subsequent sections.

Lemma 1. Suppose that $d_1, \ldots, d_n$ are a given set of n linearly independent directions. Given x_1, let $x_2, x_3, \ldots, x_{n+1}$ be the sequence of points generated by

$$x_{i+1} = x_i + \lambda_i d_i \qquad \lambda_i \neq 0$$

Suppose further that for some $t \leq n$, the gradients $g_1, \ldots, g_t$ are linearly independent but g_{t+1} is linearly dependent upon $g_1, \ldots, g_t$. Then the minimum x_{min} of $\psi(x)$ lies in the affine space

$$V = \{z : z = x_1 + \sum_{k=1}^{t} \alpha_k d_k \quad \text{for some} \quad \alpha_k \in \mathscr{R}\} \tag{3.1}$$

Proof. See Appendix 1.

Lemma 2. *Suppose that* $d_1, \ldots, d_n$ *and* $x_1, \ldots, x_{n+1}$ are given as in Lemma 1. Also let $\hat{x}_1, \ldots, \hat{x}_{n+1}$ be the sequence of points generated by exact line searches along $d_1, \ldots, d_n$ starting from x_1, i.e., $\hat{x}_1 = x_1$ and $\hat{x}_{i+1} = \hat{x}_i + \hat{\lambda}_i d_i$ $\hat{\lambda}_i$ being chosen to minimize $\psi(x)$ along d_i, starting from $\hat{x}_i$. Finally, let $\tilde{x}_{i+1} = x_i + \tilde{\lambda}_i d_i$ where $\tilde{\lambda}$ is chosen to minimize $\psi(x)$ along d_i starting from x_i.

Define e_i and ε_i by $e_i \triangleq \hat{x}_i - x_i$, $\varepsilon_i \triangleq \tilde{\lambda}_i - \lambda_i$. (3.2)

Then

$$e_{i+1} = e_i - \left(\frac{d_i^T A e_i}{d_i^T A d_i}\right) d_i + \varepsilon_i d_i \tag{3.3}$$

and

$$\varepsilon_i = -\lambda_i \, (g_{i+1}^T \, d_i / y_i^T \, d_i) \ . \tag{3.4}$$

When $d_1, \ldots, d_n$ are mutually conjugate, then

$$e_{i+1} = \sum_{j=1}^{i} \varepsilon_j d_j \quad \text{and} \quad \hat{x}_{n+1} = x_{min} \tag{3.5}$$

Proof. See Nazareth [6].

4. Basic Relations

Our results can be proved in different ways. We shall use the matrix formulation of Nazareth [15].

Suppose that the linearly independent search directions of Lemma 1 are not presented beforehand but instead developed by some algorithm, so that d_{j+1} lies in the subspace spanned by $-g_{j+1}$ and $d_1, \ldots, d_j$. Also $d_1 = -g_1$. This can be stated alternatively as d_{j+1} lies in the subspace spanned by $g_1, \ldots, g_{j+1}$. Linearly independent directions $d_1, \ldots, d_{j+1}$ can be developed provided $g_1, \ldots, g_{j+1}$ are linearly independent. By Lemma 1, if g_{t+1} becomes linearly dependent upon $g_1, \ldots, g_t$, then x_{min} lies in the affine space V (3.1). Let $D \triangleq (d_1, \ldots, d_t)$ and $G \triangleq (g_1, \ldots, g_t)$ be $n \times t$ matrices. Then $-G = DR$ where R is a $t \times t$ upper triangular matrix. We shall take $r_{ii} = 1$.

Assume further that the algorithm develops mutually conjugate directions. By Lemma 2, the step from x_{t+1} to x_{min} is determined by (3.5). Also from conjugacy $D^T AD = \alpha$, where α is a nonsingular $t \times t$ diagonal matrix.

Finally, since we are dealing with a quadratic, $g_{i+1} - g_i = A(x_{i+1} - x_i) = A\, d_i \lambda_i$, $i = 1, 2, \ldots, t$. Since $g_{t+1} = \sum_{j=1}^{t} \mu_j g_j$ for suitable μ_i, we can write these relations in the form $AD\lambda = GH$, where H is a $t \times t$ matrix of the form

$$H = \begin{bmatrix} -1 & & & & \mu_1 & \\ 1 & -1 & & & & \\ & 1 & \ddots & & & \\ & & \ddots & -1 & \mu_{t-1} & \\ & & & 1 & \mu_t & -1 \end{bmatrix} \tag{4.1}$$

and λ is the $t \times t$ diagonal matrix $\text{diag}(\lambda_1, \ldots, \lambda_t)$. In summary

$$\left.\begin{aligned} -G &= DR \\ Y \triangleq AD\lambda &= GH \\ D^T AD &= \alpha \end{aligned}\right\} \tag{4.2}$$

5. Implications

Lemma 3. Matrices G, D, R, H, λ and α which satisfy (4.2) have the following properties:

(i) RH and Y^TY are tridiagonal.

(ii) $r_{ij} = r_{ik}$, $i + 1 < j < k \leq n$, where r_{ij} denotes the i,j-th element of R.

Proof.

(i) $AD\lambda = GH$

$$AD\lambda = -D(RH) \tag{4.3}$$

using $-G = DR$.

$D^TAD\lambda = -(D^TD)(RH)$

premultiplying by D^T.

$$\text{Thus } (D^TD)^{-1} = (RH)(\alpha\lambda)^{-1} \tag{4.4}$$

using $D^TAD = \alpha$ and the non-singularity of α and λ.

But $(D^TD)^{-1}$ is symmetric and $(RH)(\alpha\lambda)^{-1}$ is upper Hessenberg.

Thus $(RH)(\alpha\lambda)^{-1}$ is tridiagonal. Since $(\alpha\lambda)$ is diagonal, this implies that (RH) is a tridiagonal matrix.

From (4.3) and (4.4), $Y^TY = (DRH)^T(DRH) = (RH)^T(D^TD)(RH)$

$$= -(RH)^T(\alpha\lambda)$$

thus Y^TY is also tridiagonal.

(ii) Using the special form of H given by (4.1) we see that

$$(RH)_{ij} = r_{ij} - r_{i,j+1} \quad \text{for} \quad i + 1 < j < n$$

$$= 0 \quad \text{using (1) above.}$$

Thus

$$r_{ij} = r_{ik} \quad \text{for} \quad i + 1 < j < k \leq n \ .$$

$\square$

Finally we note that relations $-G = DR$ and $D^T AD$ are equivalent to Gram-Schmidt orthogonalization of the gradients G in the inner product defined by A. It is easy to see that

$$r_{ij} = -y_i^T g_j / y_i^T d_i \tag{4.5}$$

N.B. The results of Sections 4 and 5 can be proved in a more conventional matter by showing that $y_i^T y_j = 0$ for $j \geq i + 1$ and using induction to prove (ii) above.

6. Algorithms

We have seen that the matrix R above is of the special form

$$R = \begin{bmatrix} 1 & r_{12} & \alpha & \alpha & \alpha & \cdots & \alpha \\ & 1 & r_{23} & \beta & \beta & \cdots & \beta \\ & & 1 & r_{34} & \gamma & \cdots & \gamma \\ & & & & \cdot & & \cdot \\ & & & & & \cdot & \cdot \\ & & & & & & \cdot \\ & & & & & & 1 \end{bmatrix} \tag{6.1}$$

Where elements denoted by the same letter of the Greek alphabet are equal. When we impose the additional requirement that all line searches be exact, all such elements are zero. The conjugate gradient method then gives search directions d_{j+1} as

$$d_1 = -g_1$$

$$d_{j+1} = -g_{j+1} + \left(\frac{y_j^T g_{j+1}}{y_j^T d_j}\right) d_j \tag{6.2}$$

and exact linear searches imply that $x_{n+1} = x_{min}$. (There are a number of expressions equivalent to (6.2) for quadratics.)

Our results show that with inexact line searches we get a very natural extension of the conjugate gradient method. The main idea is based upon Lemma 3 which shows that not all the coefficients of the Gram-Schmidt process have to be computed at every iteration. At the j + 1'th step, when computing d_{j+1}, the coefficients for $d_1, \ldots, d_{j-2}$ are already known. Thus only two new coefficients have to be computed and only two previous search directions stored. The contribution of components along $d_1, \ldots, d_{j-2}$ to the new direction d_{j+1} can be accumulated in a single vector c_j. Similarly the connection e_j to the current iterate can also be accumulated in a single vector as described in Lemma 2. This then suggests the following algorithm

$$\left.\begin{aligned}
& d_1 = -g_1 \\
& p_{j+1} = -g_{j+1} + \left(\frac{y_j^T g_{j+1}}{y_j^T d_j}\right) d_j \\
& d_{j+1} = p_{j+1} + c_j \\
& \text{where} \\
& \quad c_j = c_{j-1} + \left(\frac{y_{j-1}^T g_{j+1}}{y_{j-1}^T d_{j-1}}\right) d_{j-1} \quad \text{for } j>1 \text{ with } c_1 = 0 \\
& \quad e_j = e_{j-1} + \varepsilon_j d_j \quad \text{for } j \geq 1 \text{ with } e_0 = 0 \\
& \text{and} \\
& x_{min} = x_{n+1} + e_n \,.
\end{aligned}\right\} \qquad (6.3)$$

For arbitrary functions the linear search can ensure that x_{j+1} is chosen to satisfy $g_{j+1}^T p_{j+1} < 0$. If $g_{j+1}^T d_{j+1} \geq 0$ then the correction term can be dropped and the algorithm restarted. The algorithm has some flavor of the method proposed by Dixon [10], but it is a different formulation and will behave differently on non-quadratics.

A potential disadvantage of the above algorithm is that the recurrence relations (6.3) require that $d_1 = -g_1$. It is possible to extend the results to permit an arbitrary starting direction, in an analogous way to the extension of the conjugate gradient method with exact line searches to this case, as described by Beale [11]. This is still not entirely satisfactory, since it means defining each search direction in terms of what may be a very out of date starting direction. The three term recurrence relation, see Nazareth [6], does not require that $d_1 = -g_1$. This method has recently been combined with the conjugate gradient method in a hybrid implementation and the numerical results have been encouraging, Gill and Murray [19].

References

[1] Murtagh, B.A. and M.A. Saunders (1976), "Nonlinear Programming for Large Sparse Systems," Technical Report SOL 76-15, Department of Operations Research, Stanford University, Stanford, California.

[2] Hestenes, M.R. and E. Stiefel (1952), "Methods of Conjugate Gradients for Solving Linear Systems," Research Journal of the National Bureau of Standards, Vol. 49, pp. 409-436.

[3] Fletcher, R. and C.M. Reeves (1964), "Function Minimization by Conjugate Gradients," Computer Journal, Vol. 7, pp. 149-154.

[4] Polak, E. (1971), Computational Methods in Optimization, Academic Press, New York.

[5] Perry, A. (1976), "A Modified Conjugate Gradient Algorithm" Discussion Paper No. 229, Center for Mathematical Studies in Economics and Management Sciences, Northwestern University, Evanston, Illinois.

[6] Nazareth, J.L. (1977), "A Conjugate Direction Algorithm Without Linear Searches," J.O.T.A., (to appear).

[7] Buckley, A.G. (1976), "A Combined Conjugate Gradient Quasi-Newton Minimization Algorithm," (manuscript).

[8] Nazareth, J.L. (1976), "A Relationship Between the BFCS and Conjugate Gradient Algorithms," Argonne National Laboratory, Applied Mathematics Division, Tech. Memo. No. 282 (rev.).

[9] Shanno, D.F. (1977), "Conjugate Gradient Methods with Inexact Searches," Management Information Systems, Tech. Report No. 22, University of Arizona, Tucson, Arizona.

[10] Dixon, L.C.W. (1975), "Conjugate Gradient Algorithms: Quadratic Termination without Linear Searches," JIMA, 15, pp. 9-18.

[11] Beale, E.M.L. (1972), "A Derivation of Conjugate Gradients," in F.A. Lootsma, ed., Numerical Methods for Non-Linear Optimization, Academic Press, London, pp. 39-43.

[12] Powell, M.J.D. (1975), "Restart Procedures for the Conjugate Gradient Method," Report No. C.S.S. 24, Atomic Energy Research Establishment, Harwell, Oxfordshire, England.

[13] Lenard, M.L. (1973), "Practical Convergence Conditions for Unconstrained Optimization," Math. Prog., 4, pp. 309-323.

[14] Kawamura, K. and R.A. Volz (1973), "On the Rate of Convergence of the Conjugate Gradient Reset Method with Inaccurate Linear Minimization," IEEE Trans. on Automatic Control, Vol. Ac-18, No. 4, pp. 360-366.

[15] Cohen, A.I. (1972), "Rate of Convergence of Several Gradient Algorithms," SIAM J. Numer. Anal., 9, pp. 248-259.

[16] Daniel, J.W. (1967), "The Conjugate Gradient Method for Linear and Non-Linear Operation Equations," SIAM J. Numer. Anal., pp. 10-26.

[17] Klessig R., and E. Polak (1972), "Efficient Implementation of the Polak-Ribiere Conjugate Gradient Algorithm," SIAM J. Control, 10, pp. 524-549.

[18] Nazareth, J.L. (1977), "Unified Approach to Unconstrained Minimization via Basic Matrix Factorizations," J. Linear Algebra and its Applications, 17, pp. 197-232.

[19] Gill, P. and W. Murray (1977), (forthcoming National Physical Laboratory Report, Teddington, England).

APPENDIX 1

Proof of Lemma 1

Suppose

$$g_{j+1} = \sum_{k=1}^{j} \mu_k g_k \ . \tag{A.1}$$

Now for a quadratic

$$g_k = g_1 + \sum_{i=1}^{(k-1)} \lambda_i \, Ad_i$$

where

$$x_{i+1} = x_i + \lambda_i \, d_i \ .$$

Thus substituting in (A.1) we have

$$g_1 + \sum_{i=1}^{j} \lambda_i \, Ad_i = \sum_{k=1}^{j} \mu_k \left(g_1 + \sum_{i=1}^{(k-1)} \lambda_i \, Ad_i\right) .$$

Since A is invertible

$$-A^{-1} g_1 - \sum_{i=1}^{j} \lambda_i \, d_i = \sum_{k=1}^{j} \mu_k \, (-A^{-1} g_1) - \sum_{k=1}^{j} \mu_k \sum_{i=1}^{(k-1)} \lambda_i \, d_i \ .$$

But $x^{(min)} - x^{(1)} = -A^{-1} g^{(1)}$ for a quadratic $\psi(x)$.

Thus

$$(1 - \sum_{k=1}^{j} \mu_k)(x_{min} - x_1) = \sum_{i=1}^{j} \lambda_i d_i - \sum_{k=1}^{j} \mu_k \sum_{i=1}^{(k-1)} \lambda_i d_i$$

$$= \sum_{i=1}^{j} (1 - \sum_{k=i+1}^{j} \mu_k) \lambda_i d_i$$

with the convention $\sum_p^q \mu_k = 0$ if $p>q$.

Now if $1 - \sum_{k=1}^j \mu_k = 0$ then since $d_1, \ldots, d_j$ are linearly independent, this would imply that $\lambda_j = 0$, which contradicts the assumption inherent in the search.

Thus

$$x_{min} - x_1 = \sum_{i=1}^{j} \delta_i d_i \tag{A.2}$$

with

$$\delta_i = \lambda_i(1 - \sum_{k=i+1}^{j} \mu_k)/(1 - \sum_{k=1}^{j} \mu_k) \tag{A.3}$$

and thus x_{min} lies in the affine space V.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER SOL 78-1	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) PROPERTIES OF CONJUGATE GRADIENT METHODS WITH INEXACT LINEAR SEARCHES		5. TYPE OF REPORT & PERIOD COVERED Technical Report
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) L. Nazareth and J. Nocedal		8. CONTRACT OR GRANT NUMBER(s) N00014-75-C-0865
9. PERFORMING ORGANIZATION NAME AND ADDRESS Department of Operations Research -- SOL Stanford University Stanford, CA 94305		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS NR-047-143
11. CONTROLLING OFFICE NAME AND ADDRESS Operations Research Program -- ONR Department of the Navy 800 N. Quincy Street, Arlington, VA 22217		12. REPORT DATE January 1978
		13. NUMBER OF PAGES 15
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) UNCLASSIFIED
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

This document has been approved for public release and sale; its distribution is unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Conjugate gradient algorithms
Inexact Linear Searches
Large Dimensionality
Nonlinear Optimization

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

We show how conjugate gradient methods can be extended in a very natural way to take account of inexact linear searches.